Sparse Signal Processing with Linear and Non-Linear Observations: A Unified Shannon Theoretic Approach

Cem Aksoylar, George Atia and Venkatesh Saligrama

Abstract

In this work we derive fundamental limits for many linear and non-linear sparse signal processing models including group testing, quantized compressive sensing, multivariate regression and observations with missing features. In general sparse signal processing problems can be characterized in terms of the following Markovian property. We are given a set of N variables $X_1, X_2, \ldots, X_N$, and there is an unknown subset $S \subset \{1, 2, \ldots, N\}$ that are *relevant* for predicting outcomes/outputs Y. In other words, when Y is conditioned on $\{X_k\}_{k\in S}$ it is conditionally independent of the other variables, $\{X_k\}_{k\notin S}$.

Our goal is to identify the set S from samples of the variables X and the associated outcomes Y. We characterize this problem as a version of the noisy channel coding theorem. Using asymptotic information theoretic analyses, we describe mutual information formulas that provide sufficient and necessary conditions on the number of samples required to successfully recover the salient variables. This mutual information expression unifies conditions for both linear and non-linear observations. We then compute sample complexity bounds based on the mutual information expressions for different settings including group testing, quantized compressive sensing, multivariate regression and observations with missing features.

1 Introduction

Recent advances in sensing and storage systems have led to the proliferation of high-dimensional data such as images, video or genomic data. Such data cannot be processed efficiently using conventional signal processing methods due to their dimensionality. However, high-dimensional data often exhibit an inherent low-dimensional structure, so they can often be represented "sparsely" in some basis or domain. The discovery of an underlying sparse structure is important in order to compress the acquired data or to develop more robust and efficient processing algorithms.

In this paper, we are concerned with the asymptotic analysis of the sample complexity in problems where we aim to identify a set of *salient* variables responsible for producing an outcome. In particular, we assume that among a set of N independent and identically distributed (i.i.d.) variables/features/covariates $X = (X_1, \ldots, X_N)$, only K variables (indexed by set S) are directly relevant to the outcome Y. We formulate this with the assumption that given $X_S = \{X_n\}_{n\in S}$, outcome Y is independent of other variables $\{X_n\}_{n\notin S}$, i.e.,

$$P(Y|X) = P(Y|X_S). \tag{1}$$

We assume we are given T sample pairs (X, Y) and the problem is to identify the set of salient variables, S, from these T samples given the knowledge of observation model $P(Y|X_S)$. Our analysis aims to establish sufficient conditions on T in order to recover the set S with an arbitrarily small error probability in terms of K, N, the observation model and other model parameters such as the signal-to-noise ratio. We limit our analysis to the i.i.d. setting in this paper for simplicity. It turns out that our analysis methods can be extended to the more general dependent setting at the cost of additional terms in our formulas that compensate for dependencies between variables.

The analysis of the sample complexity is performed by posing this identification problem as an equivalent channel coding problem. The salient set S corresponds to the message transmitted through a channel. The set S is encoded by X_S^T of length T, which is the collection of codewords X_n^T for $n \in S$, from a codebook

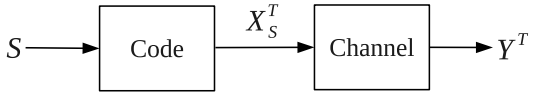

Figure 1: Channel model.

X^T. The coded message X_S^T is transmitted through a channel $P(Y|X_S)$ with output Y^T. As in channel coding, our aim is to identify which message S was transmitted given channel output Y^T and the codebook X^T.

The sufficiency and necessity results we present in this paper are analogous to the channel coding theorem for memoryless channels [10]. Our results are of the form

$$TI(X_S;Y) > \log \binom{N}{K}, \tag{2}$$

which can be interpreted as follows: The right side of the inequality is the number of bits required to represent all sets S of size K. On the left side, the mutual information term represents the uncertainty reduction on the output Y when given the input X_S, in bits per sample. This term essentially quantifies the "capacity" of the observation model $P(Y|X_S)$. Then, the total uncertainty reduction through the T samples should exceed the uncertainty of possible salient sets S, in order to reliably recover the salient set.

Sparse signal processing models analyzed in this paper have wide applicability. Below we list some examples of problems which can be formulated in the described framework.

Linear models arise naturally in array processing where the output Y is obtained as a linear (possibly noisy) transformation of some input X.

Compressive sensing (CS) [12] is a signal processing technique which aims to reconstruct a sparse signal from underdetermined linear systems. In compressed sensing, it is assumed that the output vector Y can be obtained from a K-sparse vector β through some linear transformation with basis matrix X, i.e., in the noisy case with noise W, $Y = X\beta + W$. Quantized versions of the problem are also investigated, where the channel model also includes a quantization of the output. The CS model with an example is illustrated in Figure 2. Note that, in contrast to the general CS convention, in our analysis the columns of the sensing matrix correspond to the variables X, where the support of sparse vector corresponds to the set S and the coefficients in the support are absorbed to the channel model.

Models with Missing Features [20]: Our methods also provide sample complexity bounds for sparse signal processing problems with missing features. The problem here is that the some of the variables for some of the measurements Y^T, could be missing. Specifically, we observe a $T \times N$ matrix Z^T instead of X^T, with the relation

$$Z_i^{(t)} = \begin{cases} X_i^{(t)}, & \text{w.p. } 1-\rho \\ 0, & \text{w.p. } \rho \end{cases} \qquad \forall i \in \{1,\dots,N\}, t \in \{1,\dots,T\}$$

i.e., we observe a version of the feature matrix which may have entries missing with probability ρ, independently for each entry. Interestingly our analysis shows that the sample complexity, T_{miss} for problems with missing features is related to the sample complexity, T, of the fully observed case with no missing features by the following simple expression:

$$T_{miss} = \frac{T}{1-\rho}$$

Group testing [4] is a form of compressive sensing with Boolean arithmetic. As an example, group testing has been used for medical screening to identify a set of people who have a certain disease in a large population while reducing the total number of tests. The idea is to pool blood samples from subsets of people and to test them simultaneously rather than conducting a separate blood test for each individual. In an ideal setting, the result of a test is positive if and only if the subset contains a positive sample. A significant part of

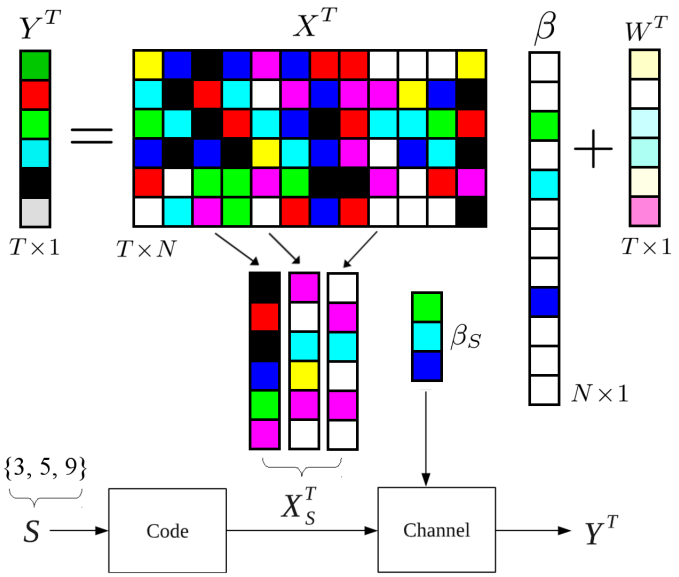

Figure 2: Compressive Sensing model example and its mapping to the channel model.

the existing research is focused on combinatorial pool design to guarantee detection using a small number of tests. Several variants of the problem exist, such as noisy group testing with different types of errors. An interesting variant is the graph-constrained group testing problem, where the salient set is the set of defective links in a graph and each test is a random walk on the graph [7]. The model can be represented graphically as in Figure 3, where X is a Boolean testing matrix and Y is the outcome vector. Again, the different columns of the testing matrix correspond to the variables X, while the defective set corresponds to set S.

Sparse Channel Estimation [9] is used for the estimation of multi-path channels characterized by sparse impulse responses. The output of the channel depends on the input time instances which correspond to the non-zero coefficients of the impulse response. In an equivalent channel model, the indices of the non-zero coefficients in the impulse response correspond to the encoded set S and the coefficients themselves are absorbed into the channel model.

Among the examples stated above, compressive sensing in particular is a fairly well-studied problem. The conditions for recovery in linear CS with measurement noise has been described and studied extensively in the literature [12, 27, 14, 26, 1, 2, 29] through the analysis of properties such as the restricted isometry property [6], as well as using information-theoretic approaches. It has been established that $T = \Omega(K \log(N/K))$ is a sufficient condition for support recovery.

Another variant of the compressive sensing problem, 1-bit CS [5] is interesting as the extreme case of CS models with quantized measurements, which are of practical importance in many real world applications. The conditions on the number of measurements have been studied for both noiseless [18] and noisy [17] models and $T = \Omega(K \log N)$ has been established as a sufficient condition for Gaussian sensing matrices.

The identification problem was formulated in a channel coding framework in [4] and in the Russian literature [24, 21, 22, 23, 13]. Sufficient and necessary conditions on the number of tests in the group testing problem with i.i.d. test assignments were derived. One main difference between the Russian literature and [4] is that, in the earlier work, the number of defective items, K, is held fixed while the number of items, N, approaches infinity. Consequently, the earlier work suggests that the number of tests must scale poly-logarithmically in N regardless of K for error probability to approach zero. In contrast, here [4] considers the fully high-dimensional setting wherein both the number of defectives as well as the number of items can approach infinity. The sufficient condition in [4] was derived based on the analysis of a Maximum Likelihood

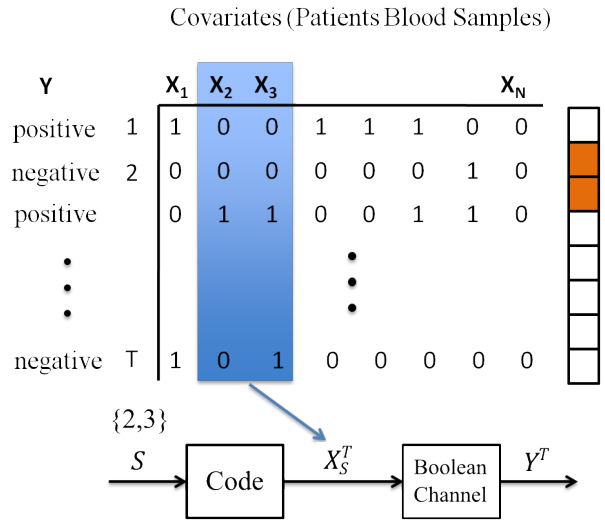

Figure 3: Group testing example and its mapping to the channel model. The codeword X is the measurement matrix, which determines whether a sample is included in a test. The result of the first test is a false positive, while the last test is a false negative.

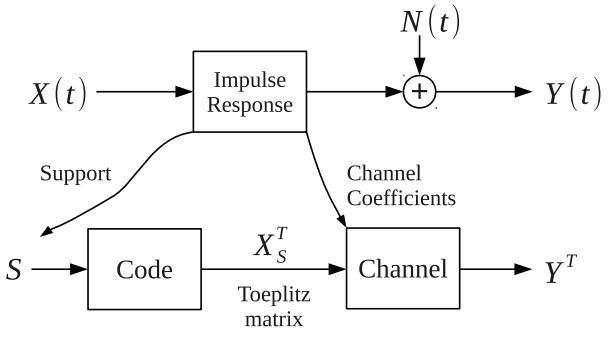

Figure 4: The channel model characterization of the sparse channel estimation problem.

(ML) decoder, while the necessary condition was derived using Fano's inequality [10]. This analysis was extended to general sparse signal processing models in [3].

In this paper, we are concerned specifically with the analysis of the problem with i.i.d. variables X, which allows analysis of a large number of important problems, such as the classical group testing or compressive sensing models. This paper presents a more thorough analysis than [3], including analysis of problems with latent variable observation models, formally extending the analysis to continuous models, presenting results for the $K = o(N)$ scaling regime and analysis of linear and 1-bit compressive sensing problems as applications.

In Section 2, we provide a general description of the problem. In Section 3, we state necessary and sufficient conditions on the number of samples required for recovery. We provide remarks for continuous models in 3.3 and we consider the scaling regime of K in 3.4. Applications are considered in Section 4, including bounds for group testing and compressive sensing models. We summarize our results in Section 5.

2 Problem Setup

We introduce our notational convention that will be used throughout the paper. We use upper case letters to denote random variables, vectors and matrices, while we use lower case letters to denote scalars, vectors and matrices, to distinguish between random quantities and their realizations. Subscripts are used for column indexing and superscripts with parentheses are used for row indexing in vectors and matrices. Subscripting with a set S implies the selection of column with indices in S. Table 1 provides a reference and further details on notation used. log is used to denote logarithm to the base 2 and natural logarithm is denoted by ln.

Let $X = (X_1, X_2, \ldots, X_N) \in \mathcal{X}^N$ denote a set of i.i.d. random variables with a joint probability distribution $Q(X)$. To simplify the expressions, we suppress the subscript indexing with random variables since the distribution function is determined solely by the number of variables indexed.

We index the different sets of size K as S_ω with index ω, so that S_ω is a set of K indices corresponding to the ω-th set of variables. Since there are N variables in total, there are $\binom{N}{K}$ such sets, hence

$$\omega \in \mathcal{I} = \left\{1, 2, \ldots \binom{N}{K}\right\}$$

Table 1: Reference for notation used

	Random quantities	**Corresponding realizations**
Variables	$X_1, \ldots, X_N$	$x_1, \ldots, x_N$
$1 \times N$ random vector	$X = (X_1, \ldots, X_N)$	$x = (x_1, \ldots, x_N)$
$1 \times \lvert S \rvert$ random vector	X_S	x_S
$T \times N$ random matrix	X^T	x^T
t-th row of X^T	$X^{(t)}$	$x^{(t)}$
n-th column of X^T	X_n^T	x_n^T
n-th element of t-th row	$X_n^{(t)}$	$x_n^{(t)}$
$T \times \lvert S \rvert$ sub-matrix	X_S^T	x_S^T
Outcome	Y	y
$T \times 1$ vector of outcomes	Y^T	y^T
t-th element of Y^T	$Y^{(t)}$	$y^{(t)}$

We let $Y \in \mathcal{Y}$ denote an observation or outcome, which depends only on a small subset of variables $S \subset \{1, \ldots, N\}$ of known cardinality $|S| = K$ where $K \ll N$. In particular, Y is conditionally independent of the variables given the subset of variables indexed by the index set S, as in (1), i.e.,

$$P(Y|X) = P(Y|X_S)$$

where $X_S = \{X_k\}_{k \in S}$ is the subset of variables indexed by the set S.

We consider a more general observation model compared to [4] and [3], where the observation model is not completely deterministic and known, but depends on a latent variable β_S. We assume β_S is independent of variables X and has a prior distribution $P(\beta_S)$. The outcomes depend on both X_S and β_S and are generated according to the model $P(Y|X_S, \beta_S)$. As an example, this latent variable corresponds to the non-zero coefficients of the K-sparse vector β in the CS framework in Section 4.1.1, or the impulse response coefficients in the sparse channel estimation framework. Note that (1) still holds in this model.

We use lower-case $p(Y|X_S)$ notation for the conditional outcome distribution given the true subset of variables averaged over the latent variable β_S. In some cases when we would like to distinguish between the outcome distribution conditioned on different sets of variables we use $p_\omega(\cdot\,|\,\cdot)$ notation, to emphasize that the conditional distribution is conditioned on the given variables, assuming the true set S is S_ω.

We observe the realizations (x^T, y^T) of T variable-outcome pairs (X^T, Y^T) with each sample realization $(x^{(t)}, y^{(t)})$ of $(X^{(t)}, Y^{(t)})$, $t = 1, 2, \ldots, T$. The variables $X^{(t)}$ are distributed i.i.d. across $t = 1, \ldots, T$. However, the outcomes $Y^{(t)}$ are independent for different t only when conditioned on β_S. Our goal is to identify the set S from the data samples and the associated outcomes (x^T, y^T), with an arbitrarily small average error probability.

We let $\hat{S}(X^T, Y^T)$ denote the estimate of the set S which is random due to the randomness in X and Y. Conditioned on a particular set S, we define the conditional error probability $P(E|S)$ as an average error probability over all possible realizations of data samples X^T and outcomes Y^T, i.e.,

$$P(E|S) = \Pr[\hat{S}(X^T, Y^T) \neq S|S] \tag{3}$$

where the randomness is over the variables X^T and the outcome Y^T. Also let $\lambda_{x^T}(S)$ denote the average error probability conditioned on a particular S and a given realization of the $T \times N$ data samples matrix x^T. Hence,

$$\lambda_{x^T}(S) = \Pr[\hat{S}(X^T, Y^T) \neq S|S, X^T = x^T] \tag{4}$$

where the randomness is over the outcome Y^T. Given (3) and (4) we have

$$P(E|S) = \sum_{x^T} \lambda_{x^T}(S) Q(x^T)$$

We further let $P(E)$ denote the average probability of error, averaged over all sets S of size K, all possible data samples X^T and outcomes Y^T, i.e.,

$$P(E) = \Pr[\hat{S}(X^T, Y^T) \neq S]$$

We assume that all sets are equally likely. So by symmetry, it is easy to see that the average error probability does not depend on the set S and we can assume without loss of generality that $\omega = 1$, i.e., S_1 is the true set.

Lastly, for any two sets S_i and S_j, we define $S_{i,j}$, $S_{i^c,j}$, and S_{i,j^c} as the overlap set, the set of indices in S_j but not in S_i, and the set of indices in S_i but not in S_j, respectively. Namely,

$$\begin{aligned} S_{i,j} &= S_i \cap S_j \quad \text{overlap} \\ S_{i^c,j} &= S_i^c \cap S_j \quad \text{in } j \text{ but not in } i \\ S_{i,j^c} &= S_i \cap S_j^c \quad \text{in } i \text{ but not in } j \end{aligned}$$

3 Conditions for Recovery

In this section we state and prove sufficient and necessary conditions for the recovery of salient set S with an arbitrarily small average error probability, for discrete variables X and for the sparsity regime where the support size K is fixed with respect to dimension N. The extensions to continuous variables and high-dimensional regime with $K = o(N)$ scaling are considered in the subsequent sections.

Central to our analysis are the following three assumptions, which we utilize in order to analyze the probability of error in recovering the salient set and to obtain sufficient and necessary conditions on sample complexity.

- **Equi-probable support:** Any set $S_\omega \subset \{1, \ldots, N\}$ with K elements is equally likely *a priori* to be the salient set. This assumption implies that we have no prior knowledge of the salient set S among $\binom{N}{K}$ sets in $\mathcal{I}$.
- **Conditional independence:** The observation/outcome Y is conditionally independent of other variables given X_S, variables with indices in S, i.e.,

  $$P(Y|X) = P(Y|X_S).$$

  This assumption follows directly from the definition of the sparse recovery problem.
- **IID variables:** The variables $X_1, \ldots, X_N$ are independent and identically distributed. While the independence assumption is not valid for all sparse recovery problems, many problems of interest can be analyzed within the i.i.d. framework, as in Section 4.

In many sparse recovery problems, we are concerned with the recovery of an underlying sparse vector β, which has a sparsity support S and coefficients β_S on the indices in the support. The observation model inherently depends on the values of these coefficients in such problems.

Defining the support coefficients as latent variables in our observation model, as stated in Section 2, such that

$$P(Y|X) = P(Y|X_S) = \int P(Y|X_S, \beta_S) P(\beta_S) \, \mathrm{d}\beta_S,$$

we are able to analyze such problems while taking their observation structure into consideration.

For instance, a simple example is the following linear observation model, where

$$Y = X^\top \beta + W = X_S^\top \beta_S + W,$$

with noise W, which exhibits such structure; along with extensions to non-linear models, where

$$Y = f(X_S^\top \beta_S) + W,$$

for a function $f : \mathbb{R} \to \mathbb{R}$.

3.1 Sufficiency

To derive the sufficiency bound for the required number of samples, we analyze the error probability of a Maximum Likelihood (ML) decoder [16]. The decoder goes through all $\binom{N}{K}$ possible sets of size K, and chooses the set S_{ω^*} for which outcome Y^T is most likely, i.e.,

$$p(Y^T|X^T_{S_{\omega^*}}) > p(Y^T|X^T_{S_\omega}); \quad \forall \omega \neq \omega^*. \tag{5}$$

An error occurs if any set other than the true set S_1 is more likely. This ML decoder is a minimum probability of error decoder assuming uniform prior on the candidate sets of variables. Note that the ML decoder requires the knowledge of the observation model $P(Y|X_S, \beta_S)$ and the prior $P(\beta_S)$. Next, we derive an upper bound

on the average error probability $P(E)$ of the ML decoder, where the average is taken over all sets, data realizations and observations.

Define the error event E_i as the event of mistaking the true set for a set which differs from the true set S_1 in exactly i variables. The probability of such an event is denoted $P(E_i)$. The event E_i implies that there exists some set which differs from the true set in i variables and is more likely to the decoder. Hence,

$$P(E_i) \leq \Pr\left[\exists \omega \neq 1 : p(Y^T|X^T_{S_\omega}) \geq p(Y^T|X^T_{S_1}), \text{ where } |S_{1^c,\omega}| = |S_{1,\omega^c}| = i, \text{ and } |S_1| = |S_\omega| = K\right] \quad (6)$$

The probability $P(E_i)$ can be written as a summation over all inputs $X^T_{S_1}$ and all outcomes Y^T

$$P(E_i) = \sum_{X^T_{S_1}} \sum_{Y^T} Q(X^T_{S_1}) p(Y^T|X^T_{S_1}) \Pr[E_i|\omega_0 = 1, X^T_{S_1}, Y^T] \quad (7)$$

where $\Pr[E_i|\omega_0 = 1, X^T_{S_1}, Y^T]$ is the probability of decoding error in exactly i variables, conditioned on the true index $\omega_0 = 1$, the realization $X^T_{S_1}$ for the set S_1, and on the sequence Y^T. This can be viewed as the error probability for a communication system with a transmitted message $\omega_0 = 1$, encoded message $X^T_{S_1}$ and received sequence Y^T. Using the union bound, the conditional error probability averaged over data realizations is upper bounded by

$$P(E|S_1) \leq \sum_{i=1}^{K} P(E_i) = \sum_{i=1}^{K} \sum_{X^T_{S_1}} \sum_{Y^T} Q(X^T_{S_1}) p(Y^T|X^T_{S_1}) \Pr[E_i|\omega_0 = 1, X^T_{S_1}, Y^T] \quad (8)$$

Next we state our main result. The following theorem provides a sufficient condition on the number of samples T for an arbitrarily small average error probability.

Theorem 3.1. *(Sufficiency). Define $\Xi_S^{\{i\}}$ as the set of tuples $(\mathcal{S}^1, \mathcal{S}^2)$ partitioning the true set S into disjoint sets $\mathcal{S}^1$ and $\mathcal{S}^2$ with cardinalities i and $K - i$, respectively, i.e.,*

$$\Xi_S^{\{i\}} = \Big\{(\mathcal{S}^1, \mathcal{S}^2) : \mathcal{S}^1 \cap \mathcal{S}^2 = \emptyset, \mathcal{S}^1 \cup \mathcal{S}^2 = S, |\mathcal{S}^1| = i, |\mathcal{S}^2| = K - i\Big\}. \quad (9)$$

If the number of samples T is such that

$$T > (1 + \epsilon) \cdot \max_{\substack{i=1,\ldots,K \\ (\mathcal{S}^1,\mathcal{S}^2) \in \Xi_S^{\{i\}}}} \frac{\log \binom{N-K}{i} \binom{K}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)}, \quad (10)$$

then, asymptotically the average error probability approaches zero, i.e.,

$$\lim_{K\to\infty} \lim_{N\to\infty} P(E) = 0,$$

where $\epsilon > 0$ is an arbitrary constant independent of N and K and $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$ is the mutual information [10] between $X_{\mathcal{S}^1}$ and $(X_{\mathcal{S}^2}, Y)$ conditioned on β_S.

Theorem 3.1 follows from a tight bound—based on characterization of error exponents as in [16]—on the error probability $P(E_i)$. We will show that the error exponent, $E_o(\rho)$, is described by:

$$E_o(\rho) = -\frac{1}{T} \log \sum_{Y^T} \sum_{X^T_{\mathcal{S}^2}} \left[\sum_{X^T_{\mathcal{S}^1}} Q(X^T_{\mathcal{S}^1}) p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1})^{\frac{1}{1+\rho}} \right]^{1+\rho} \qquad 0 \leq \rho \leq 1 \quad (11)$$

where, $(\mathcal{S}^1, \mathcal{S}^2) \in \Xi_S^{\{i\}}$, defined in (9), denoting any disjoint partitions of the set of variables S_1 with cardinalities i and $K - i$, respectively. $X^T_{\mathcal{S}^1}$ and $X^T_{\mathcal{S}^2}$ are the corresponding disjoint partitions of the $T \times K$ input $X^T_{\mathcal{S}}$ of sizes $T \times i$ and $T \times (K - i)$, respectively. We state the following lemma, which upper bounds the probability of decoding error in i variables:

Lemma 3.1. *The probability of the error event E_i defined in (7) that a set which differs from the set S_1 in exactly i variables is selected by the ML decoder (averaged over all data realizations and outcomes) is bounded from above by*

$$P(E_i) \leq 2^{-T\left(E_o(\rho)-\rho\frac{\log\binom{N-K}{i}\binom{K}{i}}{T}\right)}. \tag{12}$$

The proof of Lemma 3.1 is provided in the Appendix.

Proof of Theorem 3.1

We need to derive a sufficient condition for the error exponent of the error probability $P(E_i)$ in (12) to be positive and to drive the error probability to zero as $N \to \infty$. Specifically,

$$Tf(\rho) = TE_o(\rho) - \rho \log \binom{N-K}{i}\binom{K}{i} \to \infty \tag{13}$$

where

$$f(\rho) = E_o(\rho) - \rho\frac{\log\binom{N-K}{i}\binom{K}{i}}{T}$$

and where $E_o(\rho)$ is defined in (11).

To establish (10) we follow the argument in [16]. Note that $f(0) = 0$. Since the function $f(\rho)$ is differentiable and has a power series expansion, for a sufficiently small δ, we get by Taylor series expansion in the neighborhood of $\rho \in [0, \delta]$ that,

$$f(\rho) = f(0) + \rho\frac{df}{d\rho}\Big|_{\rho=0} + O(\rho^2)$$

But we can show that

$$\frac{\partial E_o}{\partial \rho}\Big|_{\rho=0} = \frac{1}{T}\sum_{Y^T}\sum_{X^T_{\mathcal{S}^2}}\Bigg[\sum_{X^T_{\mathcal{S}^1}} Q(X^T_{\mathcal{S}^1})p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1}) \log p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1}) - \sum_{X^T_{\mathcal{S}^1}} Q(X^T_{\mathcal{S}^1})p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1}) \log \sum_{X^T_{\mathcal{S}^1}} Q(X^T_{\mathcal{S}^1})p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1})\Bigg]$$

which simplifies to

$$\begin{aligned}\frac{\partial E_o}{\partial \rho}\Big|_{\rho=0} &= \frac{1}{T}\sum_{Y^T}\sum_{X^T_{\mathcal{S}^2}}\sum_{X^T_{\mathcal{S}^1}} Q(X^T_{\mathcal{S}^1})p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1}) \log \frac{p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1})}{\sum_{X^T_{\mathcal{S}^1}} Q(X^T_{\mathcal{S}^1})p(Y^T, X^T_{\mathcal{S}^2}|X^T_{\mathcal{S}^1})} \\ &= \frac{I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T)}{T}.\end{aligned} \tag{14}$$

Note that we can further decompose $I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T)$ using the following chain of equalities:

$$\begin{aligned}I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T) + I(\beta_S; X^T_{\mathcal{S}^1}|X^T_{\mathcal{S}^2}, Y^T) &= I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T, \beta_S) = I(X^T_{\mathcal{S}^1}; \beta_S) + I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T|\beta_S) \\ &= TI(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S),\end{aligned}$$

where the last equality is due to X and β_S being independent and (X^T, Y^T) pairs being independent over t given β_S. Therefore we have

$$\frac{\partial E_o}{\partial \rho}\Big|_{\rho=0} = \frac{I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T)}{T} = I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) - \frac{I(\beta_S; X^T_{\mathcal{S}^1}|X^T_{\mathcal{S}^2}, Y^T)}{T}. \tag{15}$$

Now assume that T satisfies

$$T > \frac{\log \binom{N-K}{i}\binom{K}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)} \tag{16}$$

which is implied by condition (10). We note that from the Lagrange form of the Taylor Series expansion (an application of the mean value theorem) we can write $E_o(\rho)$ in terms of its first derivative evaluated at zero and a remainder term, i.e.,

$$E_o(\rho) = E_o(0) + \rho E_o'(0) + \frac{\rho^2}{2} E_o''(\psi)$$

for some $\psi \in [0, \rho]$. Hence, for the choice of T in (16) and using (15) we have

$$Tf(\rho) \geq T\left(\rho \frac{\epsilon}{1+\epsilon} I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) - \rho^2 C I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) - \rho \frac{I(\beta_S; X_{\mathcal{S}^1}^T | X_{\mathcal{S}^2}^T, Y^T)}{T}\right) \tag{17}$$

where $C = \frac{|E_o''(\psi)|}{2I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)}$ which might depend on K.

A preliminary analysis of (16) reveals that $T = \Omega(\log N)$, since $\log \binom{N-K}{i}\binom{K}{i} = \Theta(i \log N)$ and $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) = I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S) \leq H(Y) = O(1)$. Also, $I(\beta_S; X_{\mathcal{S}^1}^T | X_{\mathcal{S}^2}^T, Y^T) \leq H(\beta_S)$, which is constant with respect to N since the observation model is only dependent on K variables, due to the sparsity assumption of the observation model $P(Y|X)$. So we see that

$$\frac{I(\beta_S; X_{\mathcal{S}^1}^T | X_{\mathcal{S}^2}^T, Y^T)}{T} = O\left(\frac{1}{\log N}\right)$$

which is always dominated by $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$. Therefore (17) is asymptotically equivalent to

$$Tf(\rho) \geq T\left(\rho \frac{\epsilon}{1+\epsilon} I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) - \rho^2 C I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)\right).$$

Finally, if we choose $\rho \leq \frac{\epsilon'}{C}$, where $\epsilon' = \frac{\epsilon}{1+\epsilon}$, then $f(\rho) = \delta$ for some $\delta > 0$ which does not depend on N or T. It follows that $Tf(\rho) \to \infty$ as $N \to \infty$.

We have just shown that for fixed K, $T > (1+\epsilon) \cdot \frac{\log \binom{N-K}{i}\binom{K}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)}$ is sufficient to ensure an arbitrarily small $P(E_i)$. Since the average error probability $P(E) \leq \sum_{i=1}^K P(E_i)$, it follows that for any fixed K, $\lim_{N\to\infty} \sum_{i=1}^K P(E_i) = 0$. Consequently, since this is true for any K, $\lim_{K\to\infty} \lim_{N\to\infty} \sum_{i=1}^K P(E_i) = 0$. Theorem 3.1 now follows.

□

It is important to highlight the main difference between the analysis of the error probability for the problem considered herein and the channel coding problem. In contrast to channel coding, the codewords of a candidate set and the true set are not independent since the two sets could be overlapping. To overcome this difficulty, we separate the error events E_i, $i = 1, \ldots, K$, of misclassifying the true set in i items. Then, for every i we average over realizations of ensemble of codewords for every candidate set while holding fixed the partition common to these sets and the true set of variables.

3.2 Necessity

In this section we derive lower bounds on the required number of measurements using Fano's inequality [10]. We state the following theorem:

Theorem 3.2. *For N variables and a set S_ω of K salient variables, a lower bound on the total number of measurement required to recover the set is given by*

$$T \geq \max_{\substack{i=1,\ldots,K \\ (\mathcal{S}^1, \mathcal{S}^2) \in \Xi_{S_\omega}^{\{i\}}}} \frac{\log \binom{N-K+i}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)}, \tag{18}$$

where the set $\Xi_{S_\omega}^{\{i\}}$ is the set of tuples $(\mathcal{S}^1, \mathcal{S}^2)$ partitioning the set S_ω into disjoint sets $\mathcal{S}^1$ and $\mathcal{S}^2$ with cardinalities i and $K-i$, respectively as defined in (9).

Proof. The vector of outcomes Y^T is probabilistically related to the index $\omega \in \mathcal{I} = \{1, 2, \ldots, \binom{N}{K}\}$. Suppose $K-i$ elements of the salient set are revealed to us, denoted by $\mathcal{S}^2$. From X^T and Y^T we estimate the set index ω. Let the estimate be $\hat{\omega} = g(X^T, Y^T)$. Define the probability of error

$$P_e = P(E) = \Pr[\hat{\omega} \neq \omega].$$

E is a binary random variable that takes the value 1 in case of an error i.e., if $\hat{\omega} \neq \omega$, and 0 otherwise, then using the chain rule of entropies [10] we have

$$\begin{aligned} H(E, \omega | Y^T, X^T, \mathcal{S}^2) &= H(\omega | Y^T, X^T, \mathcal{S}^2) + H(E | \omega, Y^T, X^T, \mathcal{S}^2) \\ &= H(E | Y^T, X^T, \mathcal{S}^2) + H(\omega | E, Y^T, X^T, \mathcal{S}^2). \end{aligned} \tag{19}$$

The random variable E is fully determined given X^T, Y^T, ω and $\mathcal{S}^2$. It follows that $H(E|\omega, Y^T, X^T, \mathcal{S}^2) = 0$. Since E is a binary random variable $H(E|Y^T, X^T, \mathcal{S}^2) \leq 1$. Consequently, we can bound $H(\omega|E, Y^T, X^T, \mathcal{S}^2)$ as follows,

$$\begin{aligned} H(\omega | E, Y^T, X^T, \mathcal{S}^2) &= P(E=0) H(\omega | E=0, Y^T, X^T, \mathcal{S}^2) + P(E=1) H(\omega | E=1, Y^T, X^T, \mathcal{S}^2) \\ &\leq (1 - P_e)\, 0 + P_e \log \left(\binom{N-K+i}{i} - 1 \right) \\ &\leq P_e \log \binom{N-K+i}{i}. \end{aligned} \tag{20}$$

The first inequality follows from the fact that revealing $K-i$ entries, and given that $E=1$, the conditional entropy can be upper bounded by the logarithm of the number of outcomes. From (19), we obtain the genie aided Fano's inequality

$$H(\omega | Y^T, X^T, \mathcal{S}^2) \leq 1 + P_e \log \binom{N-K+i}{i} \tag{21}$$

Note that for the left hand term, we have

$$\begin{aligned} H(\omega | Y^T, X^T, \mathcal{S}^2) &= H(\omega | \mathcal{S}^2) - I(\omega; Y^T, X^T | \mathcal{S}^2) \\ &= H(\omega | \mathcal{S}^2) - I(\omega; X^T | \mathcal{S}^2) - I(\omega; Y^T | X^T, \mathcal{S}^2) \\ &\overset{(a)}{=} H(\omega | \mathcal{S}^2) - I(\omega; Y^T | X^T, \mathcal{S}^2) \\ &\overset{(b)}{=} H(\omega | \mathcal{S}^2) - (H(Y^T | X^T, \mathcal{S}^2) - H(Y^T | X^T, \omega)) \\ &\overset{(c)}{\geq} H(\omega | \mathcal{S}^2) - (H(Y^T | X_{\mathcal{S}^2}^T) - H(Y^T | X_{S_\omega}^T)) \\ &\overset{(d)}{=} H(\omega | \mathcal{S}^2) - I(X_{\mathcal{S}^1}^T; Y^T | X_{\mathcal{S}^2}^T) \end{aligned}$$

where (a) follows from the fact that X^T is independent of $\mathcal{S}^2$ and ω; (b) follows from the fact that conditioning with respect to ω includes conditioning with respect to $\mathcal{S}^2$; (c) follows from the fact that Y^T depends on $\mathcal{S}^2$ only through $X_{\mathcal{S}^2}^T$ and similarly for the second term Y^T depends on ω only through X_S^T and finally we used the fact that conditioning reduces entropy in the first term to remove conditioning on X^T; the argument for (d) follows by definition.

From (21), it then follows that

$$H(\omega | \mathcal{S}^2) - I(X_{\mathcal{S}^1}^T; Y^T | X_{\mathcal{S}^2}^T) \leq 1 + P_e \log \binom{N-K+i}{i}$$

and since the set $\mathcal{S}^2$ of $K-i$ variables is revealed, ω is uniformly distributed over the set of indices that correspond to sets of size K containing $\mathcal{S}^2$. It follows that

$$\log \binom{N-K+i}{i} - I(X^T_{\mathcal{S}^1}; Y^T | X^T_{\mathcal{S}^2}) \leq 1 + P_e \log \binom{N-K+i}{i}.$$

Rewriting the above inequality, we have

$$P_e \geq 1 - \frac{I(X^T_{\mathcal{S}^1}; Y^T | X^T_{\mathcal{S}^2}) + 1}{\log \binom{N-K+i}{i}}. \tag{22}$$

Thus, for the probability of error to be asymptotically bounded away from zero, it is necessary that

$$\log \binom{N-K+i}{i} \leq I(X^T_{\mathcal{S}^1}; Y^T | X^T_{\mathcal{S}^2}). \tag{23}$$

Due to the independence of variables in X, we have

$$I(X^T_{\mathcal{S}^1}; Y^T | X^T_{\mathcal{S}^2}) = I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T) - I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}) = I(X^T_{\mathcal{S}^1}; X^T_{\mathcal{S}^2}, Y^T) \tag{24}$$

then, using (15), we can see that

$$T \geq \max_{i:(\mathcal{S}^1,\mathcal{S}^2)\in \Xi^{\{i\}}_{S_\omega}} \frac{\log \binom{N-K+i}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y | \beta_S) - \frac{I(\beta_S; X^T_{\mathcal{S}^1} | X^T_{\mathcal{S}^2}, Y^T)}{T}}$$

is a necessary condition for the number of samples T.

Similar to the previous proof, we note that $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y | \beta_S) = I(X_{\mathcal{S}^1}; Y | X_{\mathcal{S}^2}, \beta_S) \leq H(Y) = O(1)$ and $\log \binom{N-K+i}{i} = \Theta(i \log N)$. Therefore $T = \Omega(\log N)$ and $\frac{I(\beta_S; X^T_{\mathcal{S}^1} | X^T_{\mathcal{S}^2}, Y^T)}{T}$ is dominated by $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y | \beta_S)$, so that (18) satisfies above inequality asymptotically and is a lower bound on T.

□

Remark 3.1. *The mutual information expressions in the denominators of* (10) *and* (18) *are identical, therefore the lower bound given in Theorem 3.2 is* order-wise *tight as it matches the upper bound in Theorem 3.1.*

Remark 3.2. *Intuitively, the bounds in* (10) *and* (18) *can be explained as follows: For each i, the numerator is the number of bits required to represent all sets S_ω that differ from S in i elements. The denominator represents the information given by the subset $\mathcal{S}^2$ of $K-i$ true elements and the output variable Y about the remaining i variables $\mathcal{S}^1$. Hence, the ratio represents the number of samples needed to control i support errors and the maximization accounts for all possible support errors.*

3.3 Continuous Case

Even though the results and proof ideas that were used in sections 3.1 and 3.2 are fairly general, the proofs provided in 3.1 were stated for discrete variables and outcomes. In this section we make the necessary generalizations to extend these proofs to continuous variable and observation models. We follow the methodology in [16] and [15].

To simplify the exposition, we consider the extension to continuous variables in the special case of fixed and known β_S. In that case, $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y | \beta_S)$ reduces to $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y)$ and $E_o(\rho)$ as defined in (11) reduces to

$$E_o(\rho) = -\log \sum_Y \sum_{X_{\mathcal{S}^2}} \left[\sum_{X_{\mathcal{S}^1}} Q(X_{\mathcal{S}^1}) p(Y, X_{\mathcal{S}^2} | X_{\mathcal{S}^1})^{\frac{1}{1+\rho}} \right]^{1+\rho} \qquad 0 \leq \rho \leq 1 \tag{25}$$

with $\frac{\partial E_o(\rho)}{\partial \rho}\Big|_{\rho=0} = I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y)$, since $(X^{(t)}, Y^{(t)})$ pairs are independent across t for fixed β_S.

Assume the continuous joint variable probability density $Q(X)$ with joint cumulative density function F and the conditional probability density $p(Y = y|X = x)$ for the observation model, which is assumed to be a continuous function of both x and y.

Let $X' \in \mathcal{X}'^N$ be the random vector and $Y' \in \mathcal{Y}'$ be the random variable generated by the quantization of $X \in \mathcal{X}^N = \mathbb{R}^N$ and $Y \in \mathcal{Y} = \mathbb{R}$ respectively, where each variable in X is quantized to L values and Y quantized to J values. Let F' be the joint cumulative density function of X'. As before, let $\hat{S}(X^T, Y^T)$ be the ML decoder with continuous inputs with probability of making i errors in decoding denoted by $P(E_i)$. Let $\hat{S}(X'^T, Y'^T)$ be the ML decoder that quantizes inputs X^T and Y^T to X'^T and Y'^T, and have the corresponding probability of error $P'(E_i)$. Define

$$E_o(\rho, X', Y') = -\log \sum_{y' \in \mathcal{Y}'} \sum_{x'_{\mathcal{S}^2} \in \mathcal{X}'^{K-i}} \left[\sum_{x'_{\mathcal{S}^1} \in \mathcal{X}'^i} Q(x'_{\mathcal{S}^1}) p(y', x'_{\mathcal{S}^2} | x'_{\mathcal{S}^1})^{\frac{1}{1+\rho}} \right]^{1+\rho},$$

$$E_o(\rho, X, Y) = -\log \int_{\mathcal{Y}} \int_{\mathcal{X}^{K-i}} \left[\int_{\mathcal{X}^i} Q(x_{\mathcal{S}^1}) p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1})^{\frac{1}{1+\rho}} \, \mathrm{d}x_{\mathcal{S}^1} \right]^{1+\rho} \mathrm{d}x_{\mathcal{S}^2} \, \mathrm{d}y.$$

where the indexing denotes the random variates which the error exponents are computed with respect to.

Utilizing the results in 3.1 for the discrete models, we will show the following for the continuous model

$$P(E_i) \leq 2^{-T\left(E_o(\rho,X,Y) - \rho \frac{\log \binom{N-K}{i}\binom{K}{i}}{T}\right)}. \tag{26}$$

The rest of the proof will then follow as in the discrete case, by noting that $\frac{\partial E_o(\rho,X,Y)}{\partial \rho}\Big|_{\rho=0} = I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y)$, with the mutual information definition for continuous variables [10].

Our strategy will be the following: we will increase the number of quantization levels for Y' and X' respectively and since discrete result (12) holds for any number of quantization levels, by taking limits we will be able to show that

$$P'(E_i) \leq 2^{-T\left(E_o(\rho,X,Y) - \rho \frac{\log \binom{N-K}{i}\binom{K}{i}}{T}\right)}. \tag{27}$$

Since $\hat{S}(X^T, Y^T)$ is the minimum probability of error decoder, any upper bound for $P'(E_i)$ will also be an upper bound for $P(E_i)$, proving (26).

Assume Y is quantized with the quantization boundaries denoted by $a_1, \ldots, a_{J-1}$, with $Y' = a_j$ if $a_{j-1} < Y \leq a_j$. For convenience denote $a_0 = -\infty$ and $a_J = \infty$. Furthermore assume quantization boundaries are equally spaced, i.e. $a_j - a_{j-1} = \Delta_J$ for $2 \leq j \leq J-1$. Now we can write the following

$$E_o(\rho, X', Y') = -\log \sum_{j=1}^{J} \sum_{x'_{\mathcal{S}^2}} \left[\sum_{x'_{\mathcal{S}^1}} Q(x'_{\mathcal{S}^1}) \left(\int_{a_{j-1}}^{a_j} p(y, x'_{\mathcal{S}^2} | x'_{\mathcal{S}^1}) \, \mathrm{d}y \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{28}$$

$$= -\log \Bigg\{ \sum_{j=2}^{J-1} \Delta_J \sum_{x'_{\mathcal{S}^2}} \left[\sum_{x'_{\mathcal{S}^1}} Q(x'_{\mathcal{S}^1}) \left(\frac{\int_{a_{j-1}}^{a_j} p(y, x'_{\mathcal{S}^2} | x'_{\mathcal{S}^1}) \, \mathrm{d}y}{\Delta_J} \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{29}$$

$$+ \sum_{x'_{\mathcal{S}^2}} \left[\sum_{x'_{\mathcal{S}^1}} Q(x'_{\mathcal{S}^1}) \left(\int_{-\infty}^{a_1} p(y, x'_{\mathcal{S}^2} | x'_{\mathcal{S}^1}) \, \mathrm{d}y \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{30}$$

$$+ \sum_{x'_{\mathcal{S}^2}} \left[\sum_{x'_{\mathcal{S}^1}} Q(x'_{\mathcal{S}^1}) \left(\int_{a_{J-1}}^{\infty} p(y, x'_{\mathcal{S}^2} | x'_{\mathcal{S}^1}) \, \mathrm{d}y \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \Bigg\}. \tag{31}$$

Let $J \to \infty$ and for each J choose the sequence of quantization boundaries such that $\lim \Delta_J = 0$, $\lim a_{J-1} = \infty$, $\lim a_1 = -\infty$. Then the last two terms disappear and using the fundamental theorem of calculus, we obtain

$$\lim_{J\to\infty} E_o(\rho, X', Y') = E_o(\rho, X', Y) = -\log \int_{\mathcal{Y}} \sum_{x'_{\mathcal{S}^2}} \left[\sum_{x'_{\mathcal{S}^1}} Q(x'_{\mathcal{S}^1}) p(y, x'_{\mathcal{S}^2} | x'_{\mathcal{S}^1})^{\frac{1}{1+\rho}} \right]^{1+\rho} \mathrm{d}y. \tag{32}$$

It can also be shown that $E_o(\rho, X', Y')$ increases for finer quantizations of Y', therefore $E_o(\rho, X', Y)$ gives the smallest upper bound over $P'(E_i)$ over the quantizations of Y.

We repeat the same procedure for X. Assume each variable X_n in X is quantized with the quantization boundaries denoted by $b_1, \ldots, b_{L-1}$, with $X'_n = b_l$ if $b_{l-1} < X_n \leq b_l$. For convenience denote $b_0 = -\infty$ and $b_L = \infty$. Furthermore assume quantization boundaries are equally spaced, i.e. $b_l - b_{l-1} = \Delta_L$ for $2 \leq l \leq L-1$. Then we can write

$$E_o(\rho, X', Y) = -\log \int_Y \sum_{l=1}^{L} \left[\sum_{x'_{\mathcal{S}^1}} Q(x'_{\mathcal{S}^1}) \left(\int_{b_{l-1}}^{b_l} p(y, x_{\mathcal{S}^2} | x'_{\mathcal{S}^1}) \, \mathrm{d}x_{\mathcal{S}^2} \right)^{\frac{1}{1+\rho}} \right]^{1+\rho} \mathrm{d}y \tag{33}$$

$$= -\log \int_{\mathcal{Y}} \sum_{l=1}^{L} \left[\int_{\mathcal{X}^i} \left(\int_{b_{l-1}}^{b_l} p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1}) \, \mathrm{d}x_{\mathcal{S}^2} \right)^{\frac{1}{1+\rho}} \mathrm{d}F'(x_{\mathcal{S}^1}) \right]^{1+\rho} \mathrm{d}y \tag{34}$$

$$\begin{aligned} = -\log \int_{\mathcal{Y}} \Bigg\{ & \sum_{l=2}^{L-1} \Delta_L \left[\int_{\mathcal{X}^i} \left(\frac{\int_{b_{l-1}}^{b_l} p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1}) \, \mathrm{d}x_{\mathcal{S}^2}}{\Delta_L} \right)^{\frac{1}{1+\rho}} \mathrm{d}F'(x_{\mathcal{S}^1}) \right]^{1+\rho} \\ & + \int_{\mathcal{X}^i} \left(\int_{-\infty}^{b_1} p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1}) \, \mathrm{d}x_{\mathcal{S}^2} \right)^{\frac{1}{1+\rho}} \mathrm{d}F'(x_{\mathcal{S}^1}) \\ & + \int_{\mathcal{X}^i} \left(\int_{b_{L-1}}^{\infty} p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1}) \, \mathrm{d}x_{\mathcal{S}^2} \right)^{\frac{1}{1+\rho}} \mathrm{d}F'(x_{\mathcal{S}^1}) \Bigg\} \mathrm{d}y. \end{aligned} \tag{35}$$

where (34) follows with $F'(x_{\mathcal{S}^1})$ being the step function which represents the cumulative density function of the quantized variables $X'_{\mathcal{S}^1}$.

Let $L \to \infty$, for each L choose a set of quantization point such that $\lim \Delta_L = 0$, $\lim b_{L-1} = \infty$, $\lim b_1 = -\infty$. Again second and third terms disappear and the first sum converges to the integral over $X_{\mathcal{S}^2}$. Note that $p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1})$ is a continuous function of all its variables since it was assumed that $Q(x)$ and $p(y|x)$ were continuous. Also note that $\lim_{L\to\infty} F' = F$, which implies the weak convergence of the probability measure of X' to the probability measure of X. Given these facts, using the portmanteau theorem we obtain that $E_{F'}\left[p(Y, X_{\mathcal{S}^2} | X_{\mathcal{S}^1})\right] \to E_F\left[p(Y, X_{\mathcal{S}^2} | X_{\mathcal{S}^1})\right]$, which leads to

$$\lim_{L\to\infty} E_o(\rho, X', Y) = -\log \int_{\mathcal{Y}} \int_{\mathcal{X}^{K-i}} \left[\int_{\mathcal{X}^i} p(y, x_{\mathcal{S}^2} | x_{\mathcal{S}^1})^{\frac{1}{1+\rho}} \, \mathrm{d}F(x_{\mathcal{S}^1}) \right]^{1+\rho} \mathrm{d}x_{\mathcal{S}^2} \, \mathrm{d}y = E_o(\rho, X, Y). \tag{36}$$

This leads to the following result, completing the proof.

$$P(E_i) \leq P'(E_i) \leq \lim_{J,L\to\infty} 2^{-T\left(E_o(\rho, X', Y') - \rho \frac{\log \binom{N-K}{i}\binom{K}{i}}{T}\right)} = 2^{-T\left(E_o(\rho, X, Y) - \rho \frac{\log \binom{N-K}{i}\binom{K}{i}}{T}\right)}. \tag{37}$$

3.4 High-dimensional Case

The results in Sections 3.1 provided sufficient conditions for decoding error to go to zero asymptotically, in the case of fixed K, i.e. the given conditions ensured that the following holds true,

$$\lim_{K\to\infty}\lim_{N\to\infty} P(E) = 0. \tag{38}$$

In this section we consider the case where K scales together with N by doing the necessary analysis to find the sufficient conditions for the following expression to hold true,

$$\lim_{\substack{K=o(N)\\ N\to\infty}} P(E) = 0. \tag{39}$$

Below, we state a sufficient condition on T for a scaling regime of the mutual information expression $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$.

Theorem 3.3. *If $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$ scales with K and N such that*

$$\liminf_{K,N\to\infty} I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) > 0, \quad \forall i = 1,\dots,K,\ (\mathcal{S}^1,\mathcal{S}^2) \in \Xi_S^{\{i\}} \tag{40}$$

then,

$$T = \Omega(K\log N) \tag{41}$$

samples are sufficient for identifying set S with arbitrarily small error probability.

Proof. First, note the following,

$$P(E) \le \sum_{i=1}^{K} P(E_i) \le K\cdot\max_i P(E_i) = \max_i\ K\cdot P(E_i) \tag{42}$$

hence we have an extra $\log K$ term in the error exponent. This term did not exist in the main result since K was fixed.

For notational convenience let $I \triangleq I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$ and $E \triangleq \max_{\psi\in[0,1]}|E_o''(\psi)|$; then writing the Taylor expansion of $E_o(\rho)$ and taking into consideration (42) and that $T = \Omega(K\log N)$, we have

$$Tf(\rho) \ge T\left(\rho I - \frac{\rho^2}{2}E - \rho\frac{\log\binom{N-K}{i}\binom{K}{i}}{T} - \frac{\log K}{T} - \rho\frac{I(\beta_S; X_{\mathcal{S}^1}^T|X_{\mathcal{S}^2}^T, Y^T)}{T}\right) \tag{43}$$

and our aim is to show that the above quantity approaches infinity for some $\rho\in[0,1]$ as $K,N\to\infty$.

The condition of the theorem, $\liminf_{K,N\to\infty} I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) > 0$ for all $i = 1,\dots,K$ implies that there exists a $\gamma > 0$, independent of K and N such that

$$I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) \ge \gamma > 0 \tag{44}$$

for all i, for large enough K and N.

Then for $T = cK\log N$, we see that $\frac{\log K}{T}\to 0$ and $\frac{I(\beta_S; X_{\mathcal{S}1}^T|X_{\mathcal{S}^2}^T, Y^T)}{T} = O\left(\frac{K}{K\log N}\right)\to 0$, so that the last two terms in (43) are dominated by ρI. Therefore we obtain

$$Tf(\rho) \ge T\left(\rho I - \frac{\rho^2}{2}E - \rho\frac{1}{c}\right) \tag{45}$$

where by choosing ρ close enough to zero and leveraging the fact that E is bounded above by a constant, the second term can also be ignored, leaving us with

$$Tf(\rho) > T\rho\left(I - \frac{1}{c}\right) \tag{46}$$

which tends to infinity for a large enough choice of constant $c > 0$, as $I \ge \gamma$. □

We present a corollary with a simplified condition that can be checked to determine if $T = \Omega(K \log N)$ samples are sufficient, which directly follows from the above theorem.

Corollary 3.1. *The condition that*

$$\lim_{K,N\to\infty} I(X_S; Y|\beta_S) > 0$$

is necessary for $T = \Omega(K \log N)$ samples to be sufficient for identifying set S with arbitrarily small error probability.

4 Applications

In the sections below, we analyze and state specific results for some problems for which our necessity and sufficiency results are applicable. In the first section, we look at linear observation models and derive results for linear compressive sensing (CS) with measurement noise as a specific example, along with a multivariate regression model, where we deal with vector-valued variables and outcomes. In the second section, we analyze quantized CS and group testing (Boolean CS) as examples of non-linear observation models. Finally, we look at a general framework where some of the variables are not observed, i.e., each variable is missing with some probability.

4.1 Linear Settings

4.1.1 Compressive Sensing

Using the bounds presented in this paper for general sparse models, we derive sufficient conditions for the linear compressive sensing (CS) problem with measurement noise [12] and Gaussian sensing matrix with i.i.d. entries.

We have the following normalized model [1],

$$Y^T = X^T\beta + W^T \tag{47}$$

where X^T is the $T \times N$ sensing matrix, β is a K-sparse vector of length N with support S, W^T is the measurement noise of length T and Y^T is the observation vector of length T. In particular, we assume $X_n^{(t)}$ are Gaussian distributed random variables and the entries of the matrix are independent across rows t and columns n. Each element $X_n^{(t)}$ is zero mean and has variance $\frac{1}{T}$.

We let W^T be the observation noise of length T. We assume each element is i.i.d. with $W \sim \mathcal{N}(0, \frac{1}{SNR})$. The coefficients of the support, β_S, are i.i.d. Gaussian random variables with zero mean and variance σ^2.

In order to analyze the CS problem using the proposed sparse signal processing framework, it is important to observe how the CS model as defined above relates to the general sparse model. In the case of CS, the elements in a row of the sensing matrix correspond to variables $X_1, \ldots, X_N$ as defined in Section 2. Each row of the sensing matrix is a realization of X and rows are generated i.i.d. to form X^T. It is easy to see that assumption (1) is satisfied in both models, since each measurement $Y^{(t)}$ depends only on the linear combination of the elements $X_S^{(t)}$ that correspond to the support of β. The coefficients of this combination are given by β_S, the values of the non-zero elements of β. β_S corresponds to the latent parameter of the observation model $P(Y|X_S, \beta_S)$, which encapsulates the noise W.

For the following results, let $\alpha = \frac{i}{K}$ denote the support distortion, the ratio of misidentified elements of the support S. Note that $\frac{1}{K} \leq \alpha \leq 1$. We state the following lemma.

Lemma 4.1.

$$I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) = E\left[\frac{1}{2}\ln\left(1 + \frac{\beta_{\mathcal{S}^1}^\top \beta_{\mathcal{S}^1} SNR}{T}\right)\right].$$

where the expectation is with respect to $\beta_{\mathcal{S}^1}$ which are the coefficients in β_S corresponding to the indices in $\mathcal{S}^1$.

Proof. We write the mutual information term in the following chain of equalities, to obtain the lemma.

$$\begin{aligned} I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S) &= h(Y, X_{\mathcal{S}^2}|\beta_S) - h(Y, X_{\mathcal{S}^2}|X_{\mathcal{S}^1}, \beta_S) \\ &= h(Y|X_{\mathcal{S}^2}, \beta_S) - h(Y|X_S, \beta_S) \\ &= h\left(X_{\mathcal{S}^1}^\top \beta_{\mathcal{S}^1} + W|\beta_{\mathcal{S}^1}\right) - h(W) \\ &= E\left[\frac{1}{2}\ln\left(2\pi e\left(\frac{\beta_{\mathcal{S}^1}^\top \beta_{\mathcal{S}^1}}{T} + \frac{1}{SNR}\right)\right)\right] - \frac{1}{2}\ln\left(2\pi e\,\frac{1}{SNR}\right) \\ &= E\left[\frac{1}{2}\ln\left(1 + \frac{\beta_{\mathcal{S}^1}^\top \beta_{\mathcal{S}^1} SNR}{T}\right)\right]. \end{aligned}$$

□

A closer analysis reveals that we can effectively take the expectation inside the logarithm and replace $\beta_{\mathcal{S}^1}^\top \beta_{\mathcal{S}^1}$ with its expectation, $\alpha K\sigma^2$. Considering all values of α for exact recovery and noting that the numerator $\log\binom{N-K}{i}\binom{K}{i} = \Theta(\alpha K \log N)$, we then state the following theorem.

Theorem 4.1. *For compressive sensing with independent Gaussian sensing columns with $SNR = \Omega(\log N)$ (which is a necessary condition for recovery [1]), $T = \Omega\left(\frac{K\log N}{\sigma^2}\right)$ measurements are sufficient to recover S, the support of β, with an arbitrarily small average error probability.*

The proof is provided in the Appendix.

Remark 4.1. *For the linear CS problem, we showed that our relatively simple analysis gives us a bound asymptotically identical to the best-known bound $T = \Omega(K\log(N/K))$ [1] with an independent Gaussian sensing matrix, in the sublinear sparsity regime. Although we provided results for Gaussian distributed β_S, it is easy to obtain results for other cases such as fixed or lower bounded coefficients.*

4.1.2 Multivariate Regression

In this problem, we consider the following linear model [25], where we have a total of R linear regression problems,

$$Y_{\{r\}}^T = X_{\{r\}}^T \beta_{\{r\}} + W_{\{r\}}^T, \quad r = 1, \ldots, R$$

where for each r, $\beta_{\{r\}} \in \mathbb{R}^N$ is a K-sparse vector, $X_{\{r\}}^T \in \mathbb{R}^{T\times N}$ and $Y_{\{r\}}^T \in \mathbb{R}^T$. The relation between different tasks $r = 1, \ldots, R$ is that all $\beta_{\{r\}}$ share the same support S. This model is also called multiple linear regression or distributed compressive sensing [28] and is useful in applications such as multi-task learning [19].

It is easy to see that this problem can be formulated in our sparse recovery framework, with vector-valued outcomes Y and variables X. Namely, let $Y = (Y_{\{1\}}, \ldots, Y_{\{R\}}) \in \mathbb{R}^R$ be a vector-valued outcome, $X = (X_{\{1\}}^\top, \ldots, X_{\{R\}}^\top)^\top \in \mathbb{R}^{R\times N}$ be the collection of N vector-valued variables and $\beta = (\beta_{\{1\}}, \ldots, \beta_{\{R\}}) \in \mathbb{R}^{N\times R}$ be the collection of R sparse vectors sharing support S, making it block-sparse. This mapping is illustrated in Figure 5. Assuming independence between $X_{\{r\}}$ and support coefficients $\beta_{\{r\},S}$ across $r = 1, \ldots, R$, we have the following observation model:

$$P(Y|X) = P(Y|X_S) = \prod_{r=1}^R P(Y_{\{r\}}|X_{\{r\},S}) = \prod_{r=1}^R \int_{\mathbb{R}^K} P(Y_{\{r\}}|X_{\{r\},S}, \beta_{\{r\},S}) P(\beta_{\{r\},S})\, \mathrm{d}\beta_{\{r\},S}.$$

We present the following theorem as a straightforward extension of Theorem 3.1.

Theorem 4.2. *For the multiple regression model with R regression problems (with finite R), with independent matrices $X_{\{r\}}^T$ and support coefficients $\beta_{\{r\},S}$ for different r, the following is a sufficient condition to identify*

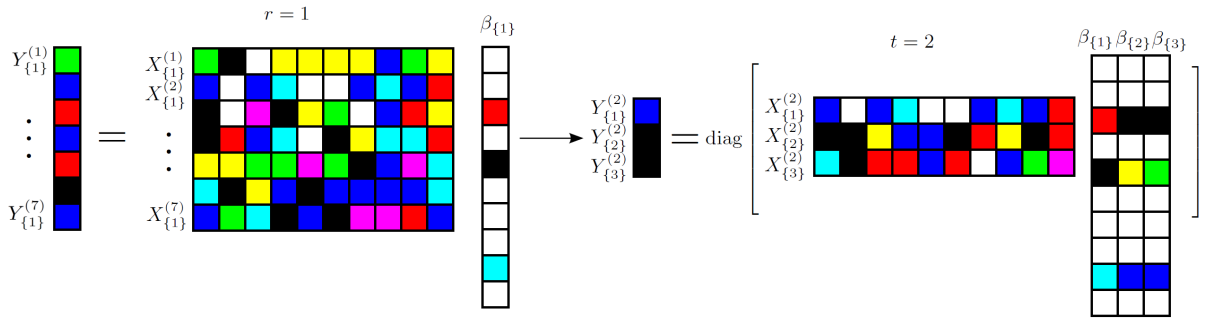

Figure 5: Mapping the multiple linear regression problem to a vector-valued outcome and variable model. On the left is the representation for a single problem $r = 1$. On the right is the corresponding vector formulation, shown for sample index $t = 2$.

the joint support set S,

$$T > (1+\epsilon) \cdot \max_{\substack{i=1,\dots,K \\ (\mathcal{S}^1,\mathcal{S}^2)\in \Xi_S^{\{i\}}}} \frac{\log \binom{N-K}{i}\binom{K}{i}}{\sum_{r=1}^{R} I(X_{\{r\},\mathcal{S}^1}; X_{\{r\},\mathcal{S}^2}, Y_{\{r\}} | \beta_{\{r\},S})}. \tag{48}$$

with an arbitrarily small average error probability.

The proof follows directly from the decomposition of $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$, due to the independence of X, Y and β_S across different r.

We present the following theorem for the specific linear model presented in Section 4.1.1, as a direct result of Theorem 4.1 and Theorem 4.2.

Theorem 4.3. *For the multiple regression model with R problems, each given by the CS framework in Section 4.1.1, with i.i.d. $X_{\{r\}}$ and $\beta_{\{r\},S}$ across R problems, $T = \Omega\left(\frac{K \log N}{R\sigma^2}\right)$ measurements are sufficient to recover joint support S with an arbitrarily small average error probability, with $SNR = \Omega(\log N)$.*

Remark 4.2. *We showed that having R problems with independent measurements and sparse vector coefficients decreases the number of measurements per problem by a factor of $1/R$. While having R such problems increases the number of measurements R-fold, the inherent uncertainty in the problem is the same since the support is shared. It is then reasonable to expect such a decrease in measurements.*

4.2 Non-linear Settings

4.2.1 1-bit Quantized Compressive Sensing

As an example of a non-linear observation model, we look at the 1-bit compressive sensing problem [5, 18, 17]. We follow the problem setup of [17]. For the 1-bit CS model, we have

$$Y^T = q(X^T\beta + W^T) \tag{49}$$

where X^T is a $T \times N$ matrix with i.i.d. standard Gaussian elements, β is a $N \times 1$ vector that is K-sparse with support S and $\beta_S = 1$. W^T is a $T \times 1$ noise vector with standard Gaussian elements. $q(\cdot)$ is a 1-bit quantizer which outputs 1 if the input is non-negative and 0 otherwise, for each element in the input vector. This setup corresponds to the $SNR = 1$ regime in [17].

To simplify the analysis and exposition, we analyze the degenerate case of $\beta \in \{0,1\}^N$, i.e. the latent variable β_S is known and equal to the vector of 1's. However, the general case where β_S are random with

a known distribution can also be analyzed using the condition given by Theorem 3.1. In order to obtain the model-specific bounds, we analyze the mutual information term $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$, which reduces to $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y)$ for this case.

Theorem 4.4. *For 1-bit CS with i.i.d. Gaussian sensing matrix and the above setup, $T = \Omega(K \log N)$ measurements are sufficient to recover S, the support of β, with an arbitrarily small average error probability.*

The proof is provided in the Appendix.

Remark 4.3. *Similar to linear CS, for 1-bit CS with noise we provided a sufficiency bound that matches [17] for an i.i.d. Gaussian sensing matrix, for the corresponding SNR regime.*

4.2.2 Group Testing - Boolean Model

In this section we look at another non-linear model, group testing. The fundamental problem of group testing can be summarized as follows. Among a population of N items, K unknown items are of interest. The collection of these K items represents the defective set. The goal is to construct a pooling design, i.e., a collection of tests, to recover the defective set while reducing the number of required tests. In this case X^T is a binary measurement matrix defining the assignment of items to tests. For the noise-free case, the outcome of the tests Y^T is deterministic. It is the Boolean sum of the codewords corresponding to the defective set S. In other words

$$Y^T = \bigvee_{i \in S} X_i^T. \tag{50}$$

Alternatively, if $R_i \in \{0, 1\}$ is an indicator function for the i-th item determining whether it belongs to the defective set (i.e. $R_i = 1$ if $i \in S$ and $R_i = 0$ otherwise), the outcome $Y^{(t)}$ of the t-th test in the noise-free case can be written as

$$Y^{(t)} = \bigvee_{i=1}^{N} X_i^{(t)} R_i \tag{51}$$

where $X_i^{(t)}$ is the t-th entry of the vector X_i^T, or equivalently, the binary entry at cell (i, t) of the measurement matrix X^T.

Theorem 4.5. *For N items and K defectives, the number of tests $T = \Omega(K \log N)$ is sufficient to identify the defective set S with an arbitrarily small average error probability. In other words, there is a constant c independent of N and K such that if $T = cK \log N$ then the average probability of error goes to zero.*

Our result also establishes upper and lower bounds on the number of tests needed for noisy versions of group testing, as well as worst-case errors. In particular, we consider testing with additive noise (leading to false alarms) and testing with dilution effects (leading to potential misses). We refer the reader to [4] for further details.

4.3 Models with Missing Features

Consider the general sparse signal processing model with independent variables, as considered in Section 3.1. However instead of trying to infer S given the features X^T and outputs Y^T, assume we observe a $T \times N$ matrix Z^T instead of X^T, with the relation

$$Z_i^{(t)} = \begin{cases} X_i^{(t)}, & \text{w.p. } 1 - \rho \\ 0, & \text{w.p. } \rho \end{cases} \qquad \forall i \in \{1, \ldots, N\}, t \in \{1, \ldots, T\}$$

i.e., we observe a version of the feature matrix which may have entries missing with probability ρ, independently for each entry. We show how the sample complexity changes related to the case where features are fully observed.

Theorem 4.6. *Assume we observe T_{miss} i.i.d. samples of missing variables Z and outcomes Y in the missing data setup. Then,*

$$T_{miss} > \frac{T_o}{1-\rho}$$

is a sufficient condition for arbitrarily small average error probability in estimating S, where T_o is the sufficiency bound given by Theorem 3.1 for fully observed variables.

Proof. We try to derive $I(Z_{\mathcal{S}^1}; Z_{\mathcal{S}^2}, Y|\beta_S)$ in terms of $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)$. To do that, we compute $H(Y|Z_S, \beta_S)$ for any set S. Define W_S as a binary vector with Bernoulli random variables such that $X_S \cdot W_S = Z_S$ with scalar multiplication. To simplify the expressions, we omit the conditioning on β_S in all entropy expressions below.

$$\begin{aligned}
H(Y|Z_S) =& H(Y, Z_S) - H(Z_S) && (52)\\
=& H(Y, Z_S, X_S, W_S) - H(X_S, W_S|Y, Z_S) - (H(Z_S, X_S, W_S) - H(X_S, W_S|Z_S)) && (53)\\
=& H(Y|Z_S, X_S, W_S) - (H(W_S|Z_S) + H(X_S|W_S, Z_S, Y)) + (H(W_S|Z_S) + H(X_S|W_S, Z_S)) && (54)\\
=& H(Y|X_S) + H(X_S|W_S, Z_S) - H(X_S|W_S, Z_S, Y) && (55)\\
=& H(Y|X_S) + \sum_{i\in S} H(X_i|W_i, Z_i) - H(X_i|W_i, Z_i, Y) && (56)\\
=& H(Y|X_S) + \sum_{i\in S} \big(\rho H(X_i|W_i = 0, Z_i) + (1-\rho)H(X_i|W_i = 1, Z_i)\big)\\
& \qquad - \big(\rho H(X_i|W_i = 0, Z_i, Y) + (1-\rho)H(X_i|W_i = 1, Z_i, Y)\big) && (57)\\
=& H(Y|X_S) + \sum_{i\in S} \rho H(X_i) - \rho H(X_i|Y) && (58)\\
=& H(Y|X_S) + \rho H(X_S) - \rho H(X_S|Y) && (59)\\
=& (1-\rho)H(Y|X_S) + \rho H(Y) && (60)
\end{aligned}$$

(52), (53) and (54) follow from the chain rule of entropy. (55) follows from the conditional independence of Y and Z_S, W_S given X_S. (56) follows from the independence of X_i, Z_i and W_i over $i \in S$. In (57) we explicitly write the conditional entropies for two values of W_i's. These expressions simplify to (58) since $X_i = Z_i$ if $W_i = 1$ and Z_i gives no information on X_i if $W_i = 0$. We group the terms over $i \in S$ to obtain (59) and again use the chain rule of entropy to obtain the final expression.

Then it simply follows that

$$\begin{aligned}
I(Z_{\mathcal{S}^1}; Z_{\mathcal{S}^2}, Y|\beta_S) &= I(Z_{\mathcal{S}^1}; Y|Z_{\mathcal{S}^2}, \beta_S) = H(Y|Z_{\mathcal{S}^2}, \beta_S) - H(Y|Z_{\mathcal{S}}, \beta_S)\\
&= (1-\rho)H(Y|X_{\mathcal{S}^2}, \beta_S) + \rho H(Y|\beta_S) - (1-\rho)H(Y|X_{\mathcal{S}}, \beta_S) - \rho H(Y|\beta_S)\\
&= (1-\rho)I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y|\beta_S)
\end{aligned}$$

□

As a special case, we obtain the following result for compressive sensing models with missing data [20]:

Theorem 4.7. *For the linear compressive sensing settings in Section 4.1.1 with $SNR = \Omega(\log N)$ and measurement matrix having missing entries with probability ρ, $T = \Omega\left(\frac{K \log N}{(1-\rho)\sigma^2}\right)$ samples are sufficient to estimate S, support of β.*

Remark 4.4. *We observe that the number of sufficient samples increases by a factor of $\frac{1}{1-\rho}$ for missing probability ρ. This example highlights the flexibility of our results due to the mutual information characterization of the model; it is easy to compute new bounds for variations of any model due to this flexibility and obtain results for very general models such as this one.*

5 Conclusions

In this paper, we stated the results of our information-theoretic analysis of the sample complexity for salient variable identification in sparse signal processing models. We characterized sufficient and necessary conditions on the number of samples, for the case of i.i.d. variables.

The results we obtain for necessary and sufficient number of samples are fairly general and applicable to a wide range of problems. The bounds only require the computation of simple mutual information expressions and characterize the trade-offs between parameters of the observation model such as SNR, number of variables N and the number of salient variables K. We provided examples of signal processing problems where such a framework applies and derived results for specific cases in Section 4.

6 Appendix

Proof of Lemma 3.1

For expositional clarity, we show the following weaker bound:

$$P(E_i) \leq 2^{-T\left(E_o(\rho) - \frac{\log\binom{N-K}{i}\binom{K}{i}}{T}\right)}. \tag{A.1}$$

Note that the main difference between the above equation and Lemma 3.1 is the missing ρ term multiplying the binomial expression. The main result follows along the same lines and we refer the reader to [4] for further details.

To prove this weaker result we denote by $\mathcal{A}_i$ the set of indices corresponding to sets of K variables that differ from the true set S_1 in exactly i variables, i.e.,

$$\mathcal{A}_i = \{\omega \in \mathcal{I} : |S_{1^c,\omega}| = i, |S_\omega| = K\} \tag{A.2}$$

We can establish that,

$$\begin{aligned}\Pr[E_i|\omega_0 = 1, X_{S_1}^T, Y^T] &\leq \sum_{\omega\in\mathcal{A}_i}\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T)\frac{p_\omega(Y^T, X_{S_{1,\omega}}^T|X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T|X_{S_{1,\omega^c}}^T)^s} \qquad (A.3)\\ &= \sum_{S_{1,\omega}}\sum_{S_{1^c,\omega}}\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T)\frac{p_\omega(Y^T, X_{S_{1,\omega}}^T|X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T|X_{S_{1,\omega^c}}^T)^s}.\end{aligned}$$

Inequality (A.3) is established separately in the Appendix. It follows that,

$$\begin{aligned}\Pr[E_i|\omega_0 = 1, X_{S_1}^T, Y^T] &\leq \left(\sum_{S_{1,\omega}}\sum_{S_{1^c,\omega}}\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T)\frac{p_\omega(Y^T, X_{S_{1,\omega}}^T|X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T|X_{S_{1,\omega^c}}^T)^s}\right)^\rho \qquad (A.4)\\ &\leq \left(\sum_{S_{1,\omega}}\binom{N-K}{i}\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T)\frac{p_\omega(Y^T, X_{S_{1,\omega}}^T|X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T|X_{S_{1,\omega^c}}^T)^s}\right)^\rho \qquad (A.5)\\ &\leq \binom{N-K}{i}\sum_{S_{1,\omega}}\left(\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T)\frac{p_\omega(Y^T, X_{S_{1,\omega}}^T|X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T|X_{S_{1,\omega^c}}^T)^s}\right)^\rho, \ \forall s>0, \ 0\leq\rho\leq 1. \qquad (A.6)\end{aligned}$$

Inequality (A.4) follows from the fact that $\Pr[E_i|\omega_0 = 1, X_{S_1}^T, Y^T] \leq 1$. Consequently, if U is an upperbound of this probability then it follows that, $\Pr[E_i|\omega_0 = 1, X_{S_1}^T, Y^T] \leq U^\rho$ for $\rho \in [0, 1]$. Inequality (A.5) follows from symmetry, namely, the inner summation is only dependent on the values of $X_{S_{1^c,\omega}}^T$ and not on the items in the set $S_{1^c,\omega}$. There are exactly $\binom{N-K}{i}$ possible sets $S_{1^c,\omega}$ hence the binomial expression. Note that the sum over $S_{1,\omega}$ cannot be further simplified. This is due to the fact that $X_{S_{1,\omega}}^T$ is already specified since we have conditioned on $X_{S_1}^T$. Since $X_{S_1}^T$ is fixed, the inner sum need not be equal for all sets $S_{1,\omega}, \omega \in \mathcal{A}_i$. Finally, (A.6) follows from standard observation that sum of positive numbers raised to ρ-th power for $\rho < 1$ is smaller than the sum of the ρ-th power of each number.

We now substitute for the conditional error probability derived above and follow the steps below:

$$P(E_i) = \sum_{X_{S_1}^T} \sum_{Y^T} p_1(X_{S_1}^T, Y^T) \Pr[E_i|\omega_0 = 1, X_{S_1}^T, Y^T]$$

$$\leq \binom{N-K}{i} \sum_{S_{1,\omega}} \sum_{Y^T} \sum_{X_{S_1}^T} p_1(X_{S_1}^T, Y^T) \left(\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T) \frac{p_\omega(Y^T, X_{S_{1,\omega}}^T | X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T | X_{S_{1,\omega^c}}^T)^s} \right)^\rho$$

Due to symmetry the summation over sets $S_{1,\omega}$ does not depend on ω. Since there are $\binom{K}{K-i}$ sets $S_{1,\omega}$ we get,

$$P(E_i) \leq \binom{N-K}{i} \binom{K}{i} \sum_{Y^T} \sum_{X_{S_1}^T} p_1(X_{S_1}^T, Y^T) \left(\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T) \frac{p_\omega(Y^T, X_{S_{1,\omega}}^T | X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T | X_{S_{1,\omega^c}}^T)^s} \right)^\rho$$

$$\leq \binom{N-K}{i} \binom{K}{i} \sum_{Y^T} \sum_{X_{S_{1,\omega^c}}^T} \sum_{X_{S_{1,\omega}}^T} Q(X_{S_{1,\omega^c}}^T) p_1(X_{S_{1,\omega}}^T, Y^T \mid X_{S_{1,\omega^c}}^T)$$

$$\left(\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T) \frac{p_\omega(Y^T, X_{S_{1,\omega}}^T | X_{S_{1^c,\omega}}^T)^s}{p_1(Y^T, X_{S_{1,\omega}}^T | X_{S_{1,\omega^c}}^T)^s} \right)^\rho$$

$$= \binom{N-K}{i} \binom{K}{i} \sum_{Y^T} \sum_{X_{S_{1,\omega^c}}^T} \sum_{X_{S_{1,\omega}}^T} Q(X_{S_{1,\omega^c}}^T) p_1^{1-s\rho}(X_{S_{1,\omega}}^T, Y^T \mid X_{S_{1,\omega^c}}^T)$$

$$\left(\sum_{X_{S_{1^c,\omega}}^T} Q(X_{S_{1^c,\omega}}^T) p_\omega(Y^T, X_{S_{1,\omega}}^T | X_{S_{1^c,\omega}}^T)^s \right)^\rho$$

$$= \binom{N-K}{i} \binom{K}{i} \sum_{Y^T} \sum_{X_{S_{1,\omega}}^T} \left(\sum_{X_{S_{1,\omega^c}}^T} Q(X_{S_{1,\omega^c}}^T) p_1^{1/(1+\rho)}(X_{S_{1,\omega}}^T, Y^T \mid X_{S_{1,\omega^c}}^T) \right)^{1+\rho}$$

where the last step follows by noting that from symmetry $X_{S_{1^c,\omega}}^T$ is just a dummy variable and can be replaced by $X_{S_{1,\omega^c}}^T$. This establishes the weaker bound in (A.1). Further details about the proof of Lemma 3.1 can be found in [4].

Proof of Equation A.3

Let ζ_ω, $\omega \in \mathcal{A}_i$ denote the event where ω is more likely than 1. Then, from the definition of $\mathcal{A}_i$, the 2 encoded messages differ in i variables. Hence

$$\Pr[E_i|\omega_0 = 1, X_{S_1}^T, Y^T] \leq P(\bigcup_{\omega \in \mathcal{A}_i} \zeta_\omega) \leq \sum_{\omega \in \mathcal{A}_i} P(\zeta_\omega)$$

Now note that $X_{S_1}^T$ shares $(K-i)$ variables with $X_{S_\omega}^T$. Following the introduced notation, the common partition is denoted $X_{S_{1,\omega}}^T$, which is a $T \times (K-i)$ submatrix. The remaining i rows which are in $X_{S_1}^T$ but not in $X_{S_\omega}^T$ are $X_{S_{1,\omega^c}}^T$. Similarly, $X_{S_{1^c,\omega}}^T$ corresponds to variables in $X_{S_\omega}^T$ but not in $X_{S_1}^T$. In other words $X_{S_1}^T = (X_{S_{1,\omega}}^T, X_{S_{1,\omega^c}}^T)$ and $X_{S_\omega}^T = (X_{S_{1,\omega}}^T, X_{S_{1^c,\omega}}^T)$, where the notation $(F^{T\times n_1}; G^{T\times n_2})$ denotes an

$T \times (n_1 + n_2)$ matrix with a submatrix F in the first n_1 columns and G in the remaining n_2 columns. Thus,

$$\begin{aligned} P(\zeta_\omega) &= \sum_{X^T_{S_\omega} : p(Y^T|X^T_{S_\omega}) \geq p(Y^T|X^T_{S_1})} Q(X^T_{S_\omega}|X^T_{S_1}) \\ &\leq \sum_{X^T_{S_{1^c,\omega}}} Q(X^T_{S_{1^c,\omega}}) \frac{p(Y^T|X^T_{S_\omega})^s}{p(Y^T|X^T_{S_1})^s} \quad \forall s > 0,\ \forall \omega \in \mathcal{A}_i \end{aligned} \tag{A.7}$$

By independence $Q(X^T_{S_1}) = Q(X^T_{S_{1,\omega}})Q(X^T_{S_{1,\omega^c}})$. Similarly, $Q(X^T_{S_\omega}) = Q(X^T_{S_{1,\omega}})Q(X^T_{S_{1^c,\omega}})$. Since we are conditioning on a particular $X^T_{S_1}$, the partition $X^T_{S_{1,\omega}}$ is fixed in the summation in (A.7) and

$$\begin{aligned} P(\zeta_\omega) &\leq \sum_{X^T_{S_{1^c,\omega}}} Q(X^T_{S_{1^c,\omega}}) \frac{p(Y^T, X^T_{S_{1,\omega}}|X^T_{S_{1^c,\omega}})^s}{Q(X^T_{S_{1,\omega}}|X^T_{S_{1^c,\omega}})^s} \frac{Q(X^T_{S_{1,\omega}}|X^T_{S_{1,\omega^c}})^s}{p(Y^T, X^T_{S_{1,\omega}}|X^T_{S_{1,\omega^c}})^s} \\ &\leq \sum_{X^T_{S_{1^c,\omega}}} Q(X^T_{S_{1^c,\omega}}) \frac{p(Y^T, X^T_{S_{1,\omega}}|X^T_{S_{1^c,\omega}})^s}{p(Y^T, X^T_{S_{1,\omega}}|X^T_{S_{1,\omega^c}})^s} \quad \forall s > 0 \end{aligned} \tag{A.8}$$

where the second inequality follows from the independence across variables, i.e. $Q(X^T_{S_{1,\omega}}|X^T_{S_{1,\omega^c}}) = Q(X^T_{S_{1,\omega}}|X^T_{S_{1^c,\omega}}) = Q(X^T_{S_{1,\omega}})$.

Proof of Theorem 4.1

Assume $T = \Theta(K \log N)$ and $SNR = \Omega(\log N)$. Then we have $\frac{KSNR}{T} = \Omega(1)$ and therefore from Lemma 4.1,

$$I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y) = \Omega\left(E\left[\log\left(1 + \frac{\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}}{K}\right)\right]\right).$$

We will now show try to show that $E\left[\log\left(1 + \frac{\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}}{K}\right)\right] = \Theta(\alpha\sigma^2)$. Define the following sequence of random variables,

$$A_K = \frac{\log\left(1 + \frac{\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}}{K}\right)}{\alpha\sigma^2}$$

then, we will have proven our claim if we can show that $\lim_{K\to\infty} E[A_K] = c$ for some constant $c > 0$.

To show that, first note that $A_K \geq 0$, $A_K \leq A_{K+1}$ and therefore $A_K \leq A_1$, for all $K = 1, 2, \ldots$. Since

$$E[A_1] = E\left[\frac{\log(1 + \beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1})}{\alpha\sigma^2}\right] \leq \frac{\log(1 + E[\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}])}{\alpha\sigma^2} = \frac{\log(1 + \alpha\sigma^2)}{\alpha\sigma^2} < \infty \tag{A.9}$$

due to Jensen's inequality, we have shown all A_K are dominated by A_1 and A_1 has finite expectation. Therefore by the dominated convergence theorem, we have

$$\lim_{K\to\infty} E[A_K] = E\left[\lim_{K\to\infty} A_K\right].$$

But for large K, $\log\left(1 + \frac{\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}}{K}\right) \approx c\frac{\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}}{K}$ for a constant $c > 0$, which can be easily seen from the Taylor expansion of the logarithm. Again, using the dominated convergence theorem, we then have

$$\lim_{K\to\infty} E[A_K] = \lim_{K\to\infty} E\left[c\,\frac{\beta^\top_{\mathcal{S}^1}\beta_{\mathcal{S}^1}}{K\alpha\sigma^2}\right] = \lim_{K\to\infty} c\,\frac{i\sigma^2}{K\alpha\sigma^2} = c,$$

showing $I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y) = \Omega(\alpha\sigma^2)$.

Then, since $\log \binom{N-K}{i}\binom{K}{i} = \Theta(i \log N)$, we can write

$$\frac{\log \binom{N-K}{i}\binom{K}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y)} = O\left(\frac{\alpha K \log N}{\alpha\sigma^2}\right) = O\left(\frac{K \log N}{\sigma^2}\right)$$

which is satisfied by $T = \Theta\left(\frac{K \log N}{\sigma^2}\right)$, proving Theorem 4.1.

Proof of Theorem 4.4

We write the mutual information term as

$$I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y) = H(Y|X_{\mathcal{S}^2}) - H(Y|X_S)$$

where we will analyze $H(Y|X_{\mathcal{S}^2})$ and $H(Y|X_S)$ to obtain a lower bound for the mutual information expression.

Defining $Z_1 = \sum_{j\in\mathcal{S}^1} X_j$, $Z_2 = \sum_{j\in\mathcal{S}^2} X_j$ and $Z = Z_1 + Z_2$, we have $H(Y|X_{\mathcal{S}^2}) = H(Y|Z_2)$ since the quantizer input $X\beta + W$ depends only on the sum of the elements of X_S. Note that $Z_2 \sim \mathcal{N}(0, D^2)$ with $D^2 = K - i$. Now we explicitly write the conditional entropy

$$H(Y|Z_2) = \int_{-\infty}^{\infty} P_{Z_2}(z) H(Y|Z_2 = z)\,\mathrm{d}z = \int_{-\infty}^{\infty} P_{Z_2}(z)\left(p_1 \log\frac{1}{p_1} + p_0 \log\frac{1}{p_0}\right)\mathrm{d}z \tag{A.10}$$

with $p_1 \triangleq \Pr[Y = 1|Z_2 = z]$ and $p_0 \triangleq 1 - p_1 = \Pr[Y = 0|Z_2 = z]$, which can be written as

$$p_1 = \Pr\left[Z_1 + Z_2 + W \geq 0 \middle| Z_2 = z\right] = \Pr\left[Z_1 + W \geq -z\right] = \Pr\left[\mathcal{N}(0, S^2) \geq -z\right] = \mathcal{Q}\left(\frac{-z}{S}\right)$$
$$p_0 = \Pr\left[Z_1 + Z_2 + W < 0 \middle| Z_2 = z\right] = \Pr\left[Z_1 + W < -z\right] = \Pr\left[\mathcal{N}(0, S^2) < -z\right] = \mathcal{Q}\left(\frac{z}{S}\right)$$

where $S^2 = i + 1$ and the $\mathcal{Q}$ function defined as $\mathcal{Q}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}}\,\mathrm{d}\tau$.

To lower bound $H(Y|Z)$, we make use of the following inequalities for $x > 0$ [8, 11]:

$$\frac{1}{12}\,e^{-x^2} \leq \mathcal{Q}(x) \leq \frac{1}{2} e^{-\frac{x^2}{2}} \tag{A.11}$$
$$1 + \frac{x^2}{2} \leq \log(2e^{\frac{x^2}{2}}) \leq \log\frac{1}{\mathcal{Q}(x)} \leq \log 12 + \frac{x^2}{\ln 2} \tag{A.12}$$

Then we write the following chain of inequalities:

$$H(Y|Z_2) = 2\int_0^\infty P_{Z_2}(z)\left(p_1 \log\frac{1}{p_1} + p_0 \log\frac{1}{p_0}\right)\mathrm{d}z \tag{A.13}$$
$$\geq 2\int_0^\infty P_{Z_2}(z) \cdot p_0 \log\frac{1}{p_0}\,\mathrm{d}z \tag{A.14}$$
$$\geq 2\int_0^\infty \frac{1}{\sqrt{2\pi D^2}} \cdot e^{-\frac{z^2}{2D^2}} \cdot \frac{1}{12} \cdot e^{-\frac{z^2}{S^2}} \cdot \left(1 + \frac{z^2}{2S^2}\right)\mathrm{d}z \tag{A.15}$$
$$= \frac{1}{12\sqrt{2\pi}D}\int_{-\infty}^\infty e^{-A\frac{z^2}{2}} \cdot \left(1 + \frac{z^2}{2S^2}\right)\mathrm{d}z \tag{A.16}$$
$$= \frac{1}{12\sqrt{2\pi}D}\left(\frac{\sqrt{2\pi}}{\sqrt{A}} + \frac{\sqrt{\pi/2}}{A^{3/2}S^2}\right) = \frac{1}{12\sqrt{A}D} + \frac{1}{24A^{3/2}DS^2} \tag{A.17}$$

Equality (A.13) follows from the evenness of the function inside the integral and we write (A.14) by noting that $p_1 \log \frac{1}{p_1}$ and $P_{Z_2}(z)$ are non-negative. $P_{Z_2}(z)$ is expanded and the above bounds for the $\mathcal{Q}$ function are used to obtain (A.15) and (A.16) is a regrouping of terms by defining $A = \frac{1}{D^2} + \frac{2}{S^2}$ and rewriting the limits of the integral by noting that the integrand is an even function. We obtain (A.17) by evaluating the integral. For A, we have

$$A = \frac{1}{K-i} + \frac{2}{i+1} = \frac{2K-i+1}{(i+1)(K-i)}$$

and replacing A, D and S, we can then write

$$\begin{aligned} H(Y|X_{S^2}) = H(Y|Z_2) &\geq c_1 \frac{\sqrt{i+1}\sqrt{K-i}}{\sqrt{2K-i+1}\sqrt{K-i}} + c_2 \frac{(i+1)^{3/2}(K-i)^{3/2}}{(2K-i+1)^{3/2}\sqrt{K-i}(i+1)} \\ &\geq c \cdot \left(\sqrt{\alpha + \frac{1}{K}} + (1-\alpha)\sqrt{\alpha + \frac{1}{K}} \right) = \Omega\left(\alpha + \frac{1}{K}\right) \end{aligned} \tag{A.18}$$

for constants $c, c_1, c_2 > 0$.

We now analyze the second term $H(Y|X_S)$ to obtain an upper bound. Again, note that $H(Y|X_S) = H(Y|Z)$, then

$$H(Y|Z) = \int_{\infty}^{\infty} P_Z(z) H(Y|Z=z)\,\mathrm{d}z = \int_{\infty}^{\infty} P_Z(z) \left(p_1 \log \frac{1}{p_1} + p_0 \log \frac{1}{p_0} \right)$$

where this time we define $p_1 \triangleq \Pr[Y=1|Z=z]$ and $p_0 \triangleq \Pr[Y=0|Z=z]$, which can be written as

$$p_1 = \Pr[Z+W \geq z | Z=z] = \Pr[W \geq -z] = \Pr[\mathcal{N}(0,1) \geq -z] = Q(-z)$$

$$p_0 = \Pr[Z+W < z | Z=z] = \Pr[W < -z] = \Pr[\mathcal{N}(0,1) < -z] = Q(z)$$

Then, write the following chain of inequalities:

$$H(Y|Z) = 2\int_0^{\infty} P_Z(z) \left(p_1 \log \frac{1}{p_1} + (1-p_1) \log \frac{1}{1-p_1} \right) \mathrm{d}z \tag{A.19}$$

$$\leq 4\int_0^{\infty} P_Z(z) \left(p_1 \log \frac{1}{p_1} \right) \mathrm{d}z \tag{A.20}$$

$$\leq 4\int_0^{\infty} \frac{1}{\sqrt{2\pi K}} e^{-\frac{z^2}{2K}} \frac{1}{2} e^{-\frac{z^2}{2}} \left(\log 12 + \frac{z^2}{2\ln 2} \right) \mathrm{d}z \tag{A.21}$$

$$= \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} e^{-B\frac{z^2}{2}} \left(\log 12 + \frac{z^2}{2\ln 2} \right) \mathrm{d}z \tag{A.22}$$

$$= \frac{\log 12}{\sqrt{2\pi K}} \frac{\sqrt{2\pi}}{\sqrt{B}} + \frac{1}{\sqrt{2\pi K} 2\ln 2} \frac{\sqrt{2\pi}}{B^{3/2}} = \frac{\log 12}{\sqrt{BK}} + \frac{1}{2\ln 2 \sqrt{K} B^{3/2}} \tag{A.23}$$

Equality (A.19) follows from the evenness of the function inside the integral and we write (A.20) by noting that $p \log \frac{1}{p} \geq (1-p) \log \frac{1}{1-p}$ for $0 \leq p \leq \frac{1}{2}$. $P_Z(z)$ is expanded and the above bounds for the $\mathcal{Q}$ function are used to obtain (A.21) and (A.22) is a regrouping of terms by defining $B = \frac{1}{K} + 1$ and rewriting the limits of the integral by noting that the integrand is an even function. We obtain (A.17) by evaluating the integral. Replacing B, we then have

$$H(Y|X_S) = H(Y|Z) \leq c_1 \frac{1}{\sqrt{K+1}} + c_2 \frac{1}{\left(\frac{1}{K}+1\right)\sqrt{K+1}} = O\left(\frac{1}{\sqrt{K+1}}\right) \tag{A.24}$$

for constants $c_1, c_2 > 0$.

Looking at (A.18) and (A.24), we have the following:

$$I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y) = H(Y|X_{\mathcal{S}^2}) - H(Y|X_S) = \Omega(\sqrt{\alpha}). \tag{A.25}$$

Finally, since $\log \binom{N-K}{i}\binom{K}{i} = \Theta(i \log N)$, we can write

$$\frac{\log \binom{N-K}{i}\binom{K}{i}}{I(X_{\mathcal{S}^1}; X_{\mathcal{S}^2}, Y)} = O\left(\frac{\alpha K \log N}{\sqrt{\alpha}}\right) = O(\sqrt{\alpha} K \log N) = O(K \log N)$$

which is satisfied by $T = \Omega(K \log N)$, proving Theorem 4.4.

References

[1] S. Aeron, M. Zhao, and V. Saligrama. Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory*, 56(10):5111–5130, Oct. 2010.

[2] M. Akcakaya and V. Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inf. Theory*, 56(1):492–504, Jan.

[3] G. Atia and V. Saligrama. A mutual information characterization for sparse signal processing. In *Proc. of Int. Colloq. on Automata, Languages and Programming (ICALP)*, Switzerland, July 2011.

[4] G. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Trans. Inf. Theory*, 58(3), March 2012.

[5] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proc. of Conf. on Information Sciences and Systems (CISS)*, pages 16–21, March 2008.

[6] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(910):589–592, 2008.

[7] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama. Graph-constrained group testing. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, pages 1913 –1917, June 2010.

[8] M. Chiani, D. Dardari, and M. K. Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *Wireless Communications, IEEE Transactions on*, 2(4):840 – 845, July 2003.

[9] S. F. Cotter and B. D. Rao. Sparse channel estimation via matching pursuit with application to equalization. *IEEE Trans. Commun.*, 50(3):374 –377, March 2002.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: John Wiley and Sons, Inc., 1991.

[11] G. T. F. de Abreu. Supertight algebraic bounds on the Gaussian Q-function. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 948 –951, Nov. 2009.

[12] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289 –1306, April 2006.

[13] A. Dyachkov. Lectures on designing screening experiments. *Moscow State Univ.*, 2003.

[14] A. K. Fletcher, S. Rangan, and V. K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory*, 55(12):5758–5772, 2009.

[15] R. G. Gallager. Information theory. In *Mathematics of Physics and Chemistry, Vol. 2*. Van Nostrand, Princeton, NJ, USA, 1964.

[16] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.

[17] A. Gupta, R. Nowak, and B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, pages 1553–1557, June 2010.

[18] L. Jacques, J.N. Laska, P.T. Boufounos, and R.G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *arXiv pre-print, arXiv:1104.3160*, 2011.

[19] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 964–972. 2010.

[20] P.-L. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *arXiv pre-print, arXiv:1109.3714*, September 2011.

[21] M. B. Malyutov. On planning of screening experiments. In *Proceedings of 1975 IEEE-USSR Workshop on Inform. Theory*, 1976.

[22] M. B. Malyutov. The separating property of random matrices. *Mat. Zametki*, 23, 1978.

[23] M. B. Malyutov. Maximal rates of screening designs. *Probability and its Applic.*, 24, 1979.

[24] M. B. Malyutov and P. S. Mateev. Screening design for non-symmetric response function. *Mat. Zemetki*, 27:109–127, 1980.

[25] S. Negahban and M. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[26] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programs. In *Allerton Conf. on Communication, Control and Computing*, Monticello, IL, 2006.

[27] M. Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, Nice, France, June 2007.

[28] M. B. Wakin, S. Sarvotham, M. F. Duarte, D. Baron, and R. G. Baraniuk. Recovery of jointly sparse signals from few random projections. In *Proceedings of the Workshop on Neural Information Processing Systems (NIPS)*, pages 1435–1442, Vancouver, Canada, Dec. 2005.

[29] Y. Wu and S. Verdu. Optimal phase transitions in compressed sensing. *IEEE Trans. Inf. Theory*, 58(10):6241–6263, Oct.